Machine learning did not outperform conventional competing risk modeling to predict revision arthroplasty

Oosterhoff, J.H.F.; Hond, A.A.H. de; Peters, R.M.; Steenbergen, L.N. van; Sorel, J.C.; Zijlstra, W.P.; ... ; Machine Learning Consortium

Clin Orthop Relat Res (2024) 482:1472-1482
DOI 10.1097/CORR.0000000000003018

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Clinical Research

Machine Learning Did Not Outperform Conventional Competing Risk Modeling to Predict Revision Arthroplasty

Jacobien H. F. Oosterhoff MD, PhD[1,2], Anne A. H. de Hond PhD[3,4,5], Rinne M. Peters MD, PhD[6], Liza N. van Steenbergen PhD[7], Juliette C. Sorel MD[8], Wierd P. Zijlstra MD, PhD[6], Rudolf W. Poolman MD, PhD[8], David Ring MD, PhD[9], Paul C. Jutte MD, PhD[10], Gino M. M. J. Kerkhoffs MD, PhD[1], Hein Putter PhD[4], Ewout W. Steyerberg MSc, PhD[3,4], Job N. Doornberg MD, PhD[10], and the Machine Learning Consortium*

Received: 11 July 2023 / Accepted: 1 February 2024 / Published online: 12 March 2024

Abstract

Background Estimating the risk of revision after arthroplasty could inform patient and surgeon decision-making. However, there is a lack of well-performing prediction models assisting in this task, which may be due to current conventional modeling approaches such as traditional survivorship estimators (such as Kaplan-Meier) or competing risk estimators. Recent advances in machine learning survival analysis might improve decision support tools in this setting. Therefore, this study aimed to assess the performance of machine learning compared with that of conventional modeling to predict revision after arthroplasty.

*Members of the Machine Learning Consortium are listed in an Appendix at the end of this article.
The first two authors contributed equally to this manuscript.
Two of the authors (JO, AdH) certify receipt of funding in an amount of less than USD 20,000 from the Dutch Arthroplasty Register (LROI, the Netherlands).
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.
Ethical approval was not sought for this study owing to the anonymized patient data registration.
This work was performed at Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands & Leiden University Medical Center, Leiden, the Netherlands.

[1]Amsterdam UMC, University of Amsterdam, Department of Orthopedic Surgery and Sports Medicine, Amsterdam, the Netherlands

[2]Department of Engineering Systems and Services, Faculty of Technology Policy and Management, Delft University of Technology, Delft, the Netherlands

[3]Clinical AI Implementation and Research Lab, Leiden University Medical Center, Leiden, the Netherlands

[4]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

[5]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands

[6]Department of Orthopaedic Surgery, Medical Center Leeuwarden, Leeuwarden, the Netherlands

[7]Dutch Arthroplasty Register (LROI), 's-Hertogenbosch, the Netherlands

[8]Department of Orthopaedic Surgery, Leiden University Medical Centre, Leiden, the Netherlands

[9]Department of Surgery and Perioperative Care, Dell Medical School, University of Texas, Austin, TX, USA

[10]Department of Orthopaedic and Trauma Surgery, University Medical Center Groningen, University of Groningen, the Netherlands

J. H. F. Oosterhoff ✉, Delft University of Technology – Faculty Technology, Policy and Management - Department of Engineering Systems and Services, Postbus 5015, 2600 GA Delft, the Netherlands, Email: j.h.f.oosterhoff@tudelft.nl

Wolters Kluwer

Question/purpose Does machine learning perform better than traditional regression models for estimating the risk of revision for patients undergoing hip or knee arthroplasty?
Methods Eleven datasets from published studies from the Dutch Arthroplasty Register reporting on factors associated with revision or survival after partial or total knee and hip arthroplasty between 2018 and 2022 were included in our study. The 11 datasets were observational registry studies, with a sample size ranging from 3038 to 218,214 procedures. We developed a set of time-to-event models for each dataset, leading to 11 comparisons. A set of predictors (factors associated with revision surgery) was identified based on the variables that were selected in the included studies. We assessed the predictive performance of two state-of-the-art statistical time-to-event models for 1-, 2-, and 3-year follow-up: a Fine and Gray model (which models the cumulative incidence of revision) and a cause-specific Cox model (which models the hazard of revision). These were compared with a machine-learning approach (a random survival forest model, which is a decision tree–based machine-learning algorithm for time-to-event analysis). Performance was assessed according to discriminative ability (time-dependent area under the receiver operating curve), calibration (slope and intercept), and overall prediction error (scaled Brier score). Discrimination, known as the area under the receiver operating characteristic curve, measures the model's ability to distinguish patients who achieved the outcomes from those who did not and ranges from 0.5 to 1.0, with 1.0 indicating the highest discrimination score and 0.50 the lowest. Calibration plots the predicted versus the observed probabilities; a perfect plot has an intercept of 0 and a slope of 1. The Brier score calculates a composite of discrimination and calibration, with 0 indicating perfect prediction and 1 the poorest. A scaled version of the Brier score, 1 – (model Brier score/null model Brier score), can be interpreted as the amount of overall prediction error.
Results Using machine learning survivorship analysis, we found no differences between the competing risks estimator and traditional regression models for patients undergoing arthroplasty in terms of discriminative ability (patients who received a revision compared with those who did not). We found no consistent differences between the validated performance (time-dependent area under the receiver operating characteristic curve) of different modeling approaches because these values ranged between -0.04 and 0.03 across the 11 datasets (the time-dependent area under the receiver operating characteristic curve of the models across 11 datasets ranged between 0.52 to 0.68). In addition, the calibration metrics and scaled Brier scores produced comparable estimates, showing no advantage of machine learning over traditional regression models.
Conclusion Machine learning did not outperform traditional regression models.
Clinical Relevance Neither machine learning modeling nor traditional regression methods were sufficiently accurate in order to offer prognostic information when predicting revision arthroplasty. The benefit of these modeling approaches may be limited in this context.

Introduction

Various predictive modeling tools have been developed and are used for decision support in healthcare to inform patient and surgeon decision-making. In orthopaedic surgery, studies have predicted revision arthroplasty using competing risk analyses [4, 6, 12, 14, 19, 22-24, 29, 37, 39]. Revision arthroplasty typically involves partial or complete exchange of the prosthesis implanted during the initial (sometimes called primary or index) surgical procedure. In a typical survival setting, only one outcome is studied, such as revision or death. However, the cumulative incidence of revision (primary outcome) depends not only on the effect of covariates (such as age or gender) but also on patient survival, because patients who have died cannot subsequently undergo revision. Standard survival analyses (Kaplan-Meier curves) treat death simply as censored information, but this approach may overestimate revision rates [13]. Therefore, in certain settings where a competing risk (such as death) is considered likely to influence the occurrence of another event (revision in our setting), a competing risk analysis should be performed with revision as the primary outcome event and death as a competing risk.

However, there is a lack of well-performing prediction models assisting in this task, which may be owing to current conventional modeling approaches such as traditional survivorship estimators (such as Kaplan-Meier) or competing risk estimators (competing risk analyses). Recent advancements in machine learning survival analysis could improve decision support tools in this setting. However, it is unclear whether machine learning generates better risk estimates than the traditional approach, although there is some preliminary evidence. A recent study from our group compared machine learning and logistic regression algorithms to predict binary events (such as reoperation: yes or no) in nine orthopaedic trauma datasets; machine learning's benefit was shown to be limited [21]. To best of our knowledge, no study to date has compared competing risk survival models based on machine learning and traditional regression methods in multiple datasets. We therefore sought to compare the performance of machine learning survival analysis and traditional regression modeling in a competing risk setting. We analyzed 11 datasets including patients undergoing arthroplasty surgery registered in the Dutch Arthroplasty Register to answer this question: Does machine learning survival analysis with competing

risk perform better than traditional regression models for estimating the risk of revision for patients undergoing hip or knee arthroplasty?

Materials and Methods

Overview: The Survival Analysis Problem in Orthopaedic Cohorts

Survival analysis for predictive modeling of orthopaedic outcomes is used to estimate the time it takes for specific events to occur. In arthroplasty, this often involves analyzing the duration until a revision surgery after a primary procedure. Therefore, this method is commonly referred to as a time-to-event analysis, where the event of interest (such as revision) lies in the future. The objective of a survival analysis is to consider factors associated with revision and estimate the likelihood that revision may occur in the future. Factors associated with revision could be patient-specific (such as age or gender) or surgery-specific (such as the type of implant), and they are included as variables in the prediction model. This estimation can inform patients and clinicians when choosing a specific treatment option. However, patients may be lost to follow-up or may die during the follow-up period, resulting in censoring of the data, which is a fundamental challenge in survival analysis (Fig. 1). Therefore, multiple survival analysis methods exist to account for censoring (Fine and Gray and cause-specific Cox). More recently, machine-learning techniques have been applied to orthopaedic prediction modeling and can improve the performance of prediction models. In the context of survival analysis, a machine-learning technique called random survival forests shows promise, especially in situations involving complex censored data.

Our primary study goal was to estimate the model performance of machine learning compared with that of conventional modeling to estimate the likelihood of revision after arthroplasty in the presence of censored data. To achieve this, we included 11 datasets and separately compared the performance of machine learning with that of conventional modeling in those 11 datasets.

Study Design and Setting

Eligible datasets were derived from previously published studies, including patients registered in the Dutch Arthroplasty Register [9] and undergoing a partial or total knee or hip arthroplasty. The overall data completeness for primary knee and hip arthroplasties was 96% in 2014 and up to 100% in 2020 [9].

This study was conducted according to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research and the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis guidelines [7, 17].

Participants' Baseline Characteristics

We queried the Dutch Arthroplasty Register, a national registry covering all Dutch hospitals performing arthroplasties. The overall data completeness for all hip and knee arthroplasties was 99% in 2018 and 100% in 2022. Data completeness for registered hip and knee revision arthroplasties was 97% in 2018 and 100% in 2022. In total, 11 datasets from published studies from the Dutch Arthroplasty Register reported on factors associated with revision or survival after partial or total knee and hip arthroplasty between 2018 and 2022 and were therefore included in our study. We developed a set of time-to-event models for each dataset, leading to a total of 11 comparisons. All were observational registry studies that reported on factors associated with revision after partial or total knee and hip arthroplasty [4-6, 12, 15, 19, 23, 24, 29, 37, 38] (Table 1). The sample size of these datasets ranged from 3038 to 218,214 procedures. The raw datasets supplied by the Dutch Arthroplasty Register were directly derived from previous studies and contained several processing steps. This resulted in different patients and variables being available across the different datasets. We therefore chose to compare machine learning versus traditional statistics in each dataset separately with the same inclusion criteria and set of associated factors as applied by the original studies. This also allowed for a direct comparison with the results from the original studies. The baseline characteristics of the 11 included datasets can be found in the original studies [4-6, 12, 15, 19, 23, 24, 29, 37, 38].

Traditional Survival Approaches

Of the included studies, one conducted a multivariable logistic regression analysis [24], five applied Kaplan-Meier analyses [4-6, 14, 37], and 10 used multivariable Cox proportional hazard regression analyses [4-6, 12, 15, 19, 23, 29, 37, 38]. None of these methods accounted for competing risks.

Survival Approaches Accounting for Competing Risks

On the included studies, we developed a set of time-to-event models with revision as the event of interest and death as the competing risk for all included studies separately.

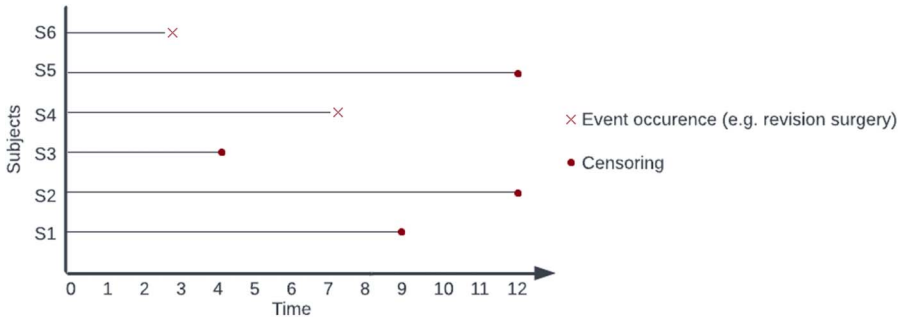

Fig. 1 This illustration shows the survival analysis.

Conventional Aalen-Johansen Curves

An Aalen-Johansen estimator is a nonparametric estimation of risks, similar to the Kaplan-Meier estimator of survival. The Aalen-Johansen curve plots the cumulative incidence function of the event of interest (revision) accounting for a competing risk (death) [1]. These curves provide insights into the probability of experiencing different types of events over time when multiple events (revision and death) are present. Aalen-Johansen curves (cumulative incidence functions) were plotted for the 11 datasets.

Conventional Fine and Gray Model

A Fine and Gray model [10] is a semiparametric method (proportional hazards model), estimating the incidence of the outcome of interest (revision) over time in the presence of a competing risk (death), thereby relating covariates to the cumulative incidence function of the event of interest (revision) [3].

Conventional Cause-specific Cox Model

A cause-specific Cox model is also a semiparametric method. It is an extension of the described Cox regression analyses. In the cause-specific Cox model, the revision risk is compared among patients who are event free and in follow-up (that is, patients who have not experienced a revision or the competing risk [death] at a particular time point) [26, 35].

Machine Learning: Random Survival Forest

The random survival forest [11] was introduced as a time-to-event extension to a random forest that can account for competing risks. Random survival forest is a machine-learning method that uses ensemble learning on many decision trees. It can work with high-dimension and complex (as well as nonlinear) data. A random survival forest shows promise, especially in situations involving complex censored data, and may be easier to interpret than other deep-learning survival models.

Data Preparation

Factors associated with revision were identified based on the original variable selection of the included studies (Table 1, Supplemental Tables 1-11; http://links.lww.com/CORR/B279). Observations where age or gender were missing were removed from the analysis. All other missing data were imputed using multivariate imputation by chained equations [32] creating 11 imputed datasets as previously applied by our group [21].

Model Development

For each of the 11 datasets, we plotted the cumulative incidence function for revision (outcome of interest) and competing risk (death) in Aalen-Johansen curves [1]. Subsequently, we compared the predictive performance of two state-of-the-art statistical time-to-event models: a Fine and Gray model and a cause-specific Cox model. These were compared with a machine-learning approach consisting of a random survival forest with competing risks [11].

The time-to-event was set at 1, 2, and 3 years of follow-up for each cohort. The imputed data were split into a training set (two-thirds of the data) and a test set. This approach was chosen over more sophisticated train designs (such as nested cross-validation) because of its computational feasibility. The hyperparameters for the random survival forest were set via fivefold cross-validation on the training data (Supplemental Table 12; http://links.lww.com/CORR/B279). The models were trained on the training data (with tuned hyperparameters) and applied to the test data.

Table 1. Characteristics of the included studies

Author	Methodology	Number of patients	Outcome	Predictors included	Study period	Main findings
Peters et al. [24]	Logistic regression	218,214	Survival; 1-year and 3-year revision	Age, gender, ASA, previous operation, smoking, BMI, Charnley	2007 to 2018	ASA and BMI were the strongest predictors for short-term revision after primary THA
Peters et al. [23]	Cox proportional hazards	209,912	Survival; revision at 5 years and 9 years	Age, gender, ASA, diagnosis, previous operation, fixation, head diameter, surgical approach, and period of surgery	2007 to 2016	A mid-term lower risk of revision was found for CoHXLPE, CoC, and Ox(HXL)PE compared with traditional MoPE-bearing surfaces
van Steenbergen et al. [38]	Cox proportional hazards	211,002	Survival; 8-year revision	Age, gender, ASA score, diagnosis (OA vs non-OA), period	2007 to 2016	Large head MoM hip arthroplasties performed significantly worse compared to non-MoM THA
van Oost et al. [37]	Kaplan-Meier, Cox proportional hazards	18,134	Survival	Age category, sex, ASA, year, diagnosis, unicondylar side, type of hospital	2007 to 2016	Higher risk of revision for partial knee replacements was seen in low absolute volume hospitals
Burger et al. [6]	Kaplan-Meier, Cox proportional hazards	19,832	Survival; 5-year revision	Age, gender, diagnosis, prior operation, bearing type, and fixation type	2007 to 2017	There is a notable risk for revision when using mobile-bearing designs for lateral UKA
Kuijpers et al. [15]	Kaplan-Meier, Cox proportional hazards	19,682	Survival; 5-year revision	Age, gender, diagnosis, ASA, surgical approach, fixation, bearing type, head size, and year	2007 to 2017	The risk of revision in patients younger than 55 years depends on surgical approach, head size, and bearing type
Bloemheuvel et al. [5]	Kaplan-Meier, Cox proportional hazards	15,922	Survival; 5-year re-revision	Gender, age, ASA, fixation	2007 to 2016	The 5-year cup re-revision rates for dual mobility cups were lower than for unipolar cups.
Bloemheuvel et al. [4]	Kaplan-Meier, Cox proportional hazards	3038	Survival; 5-year cup revision	Gender, age, diagnosis, previous operation, ASA, fixation, surgical approach, and femoral head diameter	2007 to 2016	The 5-year cup revision rates for dual mobility cup THA and unipolar cup THA were comparable
Spekenbrink-Spooren et al. [29]	Cox proportional hazards	133,841	Survival; 8-year revision	Age, gender, ASA, and previous operations	2007 to 2016	Higher mid-term revision rates of posterior stabilized TKA compared with cruciate retaining TKA

Table 1. continued

Author	Methodology	Number of patients	Outcome	Predictors included	Study period	Main findings
Moerman et al. [19]	Cox proportional hazards	30,830	Survival; 1-year revision	Gender, age, ASA, smoking BMI, approach, and stem fixation	2007 to 2017	Posterolateral approach and an uncemented hip stem have higher risks for revision surgery compared with an anterolateral approach and cemented stem
Janssen et al. [12]	Cox proportional hazards	63,354	Survival	Age, sex, diagnosis, ASA, earlier surgeries, and coating and material of stem	2007 to 2013	In THA, cementless femoral stems with a proximal shoulder are associated with early aseptic loosening when inserted through an anterior or anterolateral approach compared with a posterior approach

ASA = American Society of Anesthesiologists classification; OA = osteoarthritis; UKA = unicompartimental knee arthroplasty; CoHXLPE = ceramic-on-HXLPE; CoC = ceramic-on-ceramic; Ox(HXL)PE = oxidized-zirconium-on-(HXL)polyethylene; MoPE = metal-on-polyethylene; MOM = metal-on-metal.

Model Performance

Model performance was evaluated following recent guidance for prediction models in the presence of competing risks [36] that includes discrimination with a time-dependent area under the receiver operating curve (AUCt), calibration with a calibration slope and intercept (in line with the method by Cox [8]), and the overall prediction error with the scaled version of the Brier score [36].

Discrimination

Discrimination is a model's ability to distinguish patients who had the outcome (that is, patients who underwent revision) from those who did not [30]. It is measured using the c-index (AUC), which ranges from 0.50 to 1.0, with 1.0 indicating the highest discrimination score (most effective discrimination) and 0.50 indicating a similar chance as flipping a coin. The time-dependent c-index (AUCt) can be calculated for a single timepoint of interest (such as 2-year revision) [36].

Model performance estimates were pooled across the 11 imputed datasets via Rubin's Rules [27]. We visualized model performance comparison in a bee-swarm plot, which is a scatterplot of the differences in AUCt of each machine learning and traditional regression pair.

Calibration

Calibration reflects the difference between the likelihood of an event as a model predicts it and the actual, observed frequency of the event in question. A calibration plot plots the primary outcome's estimated and observed probabilities. A perfect calibration plot has an intercept of 0 (< 0 reflects overestimation, > 0 reflects underestimating the probability of the outcome) and a slope of 1 (the model is performing similarly in training and test sets) [31, 33]. In a small dataset, the slope is often < 1, reflecting model overfitting; probabilities are too extreme (low probability is too low, and high probability is too high) [37].

Overall Prediction Error

Overall prediction error is a composite of discrimination and calibration, and is measured using the Brier score. A Brier score of 0 indicates perfect prediction and a Brier score of 1 reflects the poorest prediction [30]. A scaled version of the Brier score, 1 – (model Brier score/null model Brier score), can be interpreted as the amount of prediction error in a null model that the prediction model explains. A 100% scaled Brier score corresponds to a perfect model, 0% to an ineffective model, and < 0% to a harmful model [36].

Software

Data preprocessing and analysis was performed using R Version 5.3 (the R Foundation), R- studio Version 1.2.1335 (R-Studio), and Python version 3.10. The following packages were used: caret, cmprsk, geepack, Hmisc, modelr, prodlim, randomForestSRC, riskRegression, survival, tidyr, tidyverse, and beeswarm. We used the following packages for Python: pandas, numpy, matplotlib, lifelines, sksurv, and sklearn.

Ethical Approval

The Dutch Arthroplasty Register database consists of anonymized patient data registration; therefore, informed consent was not necessary. Institutional research board approval was not required because of the retrospective nature of the study.

Results

We found no differences between machine learning survivorship analysis using a competing risks estimator and traditional regression models for patients undergoing arthroplasty in terms of discriminative ability (distinguishing patients who received a revision from those who did not).

On average, the differences between the validated performance (AUCt) of different modeling approaches ranged from -0.04 to 0.03 across the 11 datasets (the AUCt of the models ranged between 0.52 and 0.68 (Supplemental Table 13; http://links.lww.com/CORR/B279). There were no consistent differences between the 11 datasets; on average (the mean difference between the modeling approaches across 11 comparisons), the difference was 0.00. These findings indicate that machine learning and traditional regression models produce similar probability estimates (Fig. 2).

In addition, there were no consistent differences in calibration metrics (Supplemental Table 14; http://links.lww.com/CORR/B279) and overall prediction error (Supplemental Table 15; http://links.lww.com/CORR/B279), showing no advantage of machine learning over conventional modeling.

The cumulative incidence functions (Aalen-Johansen curves) are shown for the 11 datasets (Fig. 3). For most datasets, the absolute risk of death surpassed the risk of revision at some timepoint, which concurs with the population that was generally studied.

The results indicate that the prediction models developed using the 11 original datasets performed poorly on discrimination, calibration, and overall prediction error.

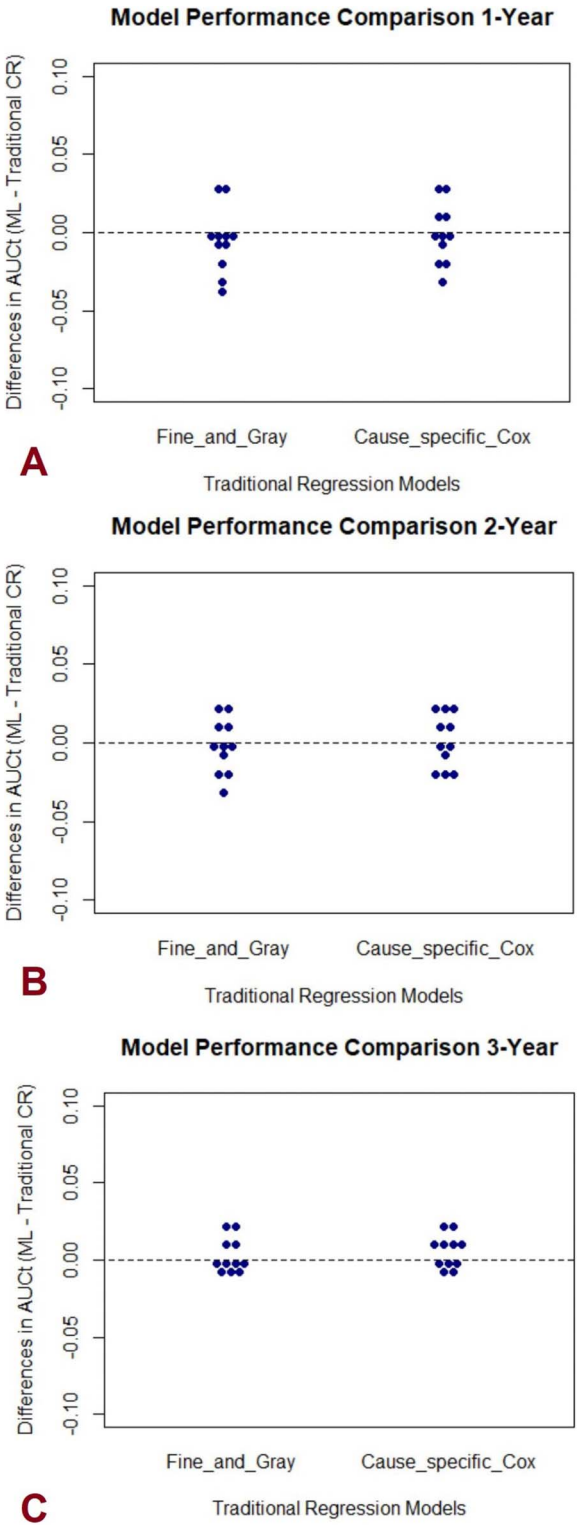

Fig. 2 Bee-swarm plots of differences in model performance AUCt (Δmachine learning – traditional regression) are shown here. (**A**) Shows a comparison of the model's performance at 1-year follow-up. (**B**) Shows a comparison of the model's performance at 2 years of follow-up. (**C**) Shows a comparison of the model's performance at 3 years of follow-up. CR = competing risk.

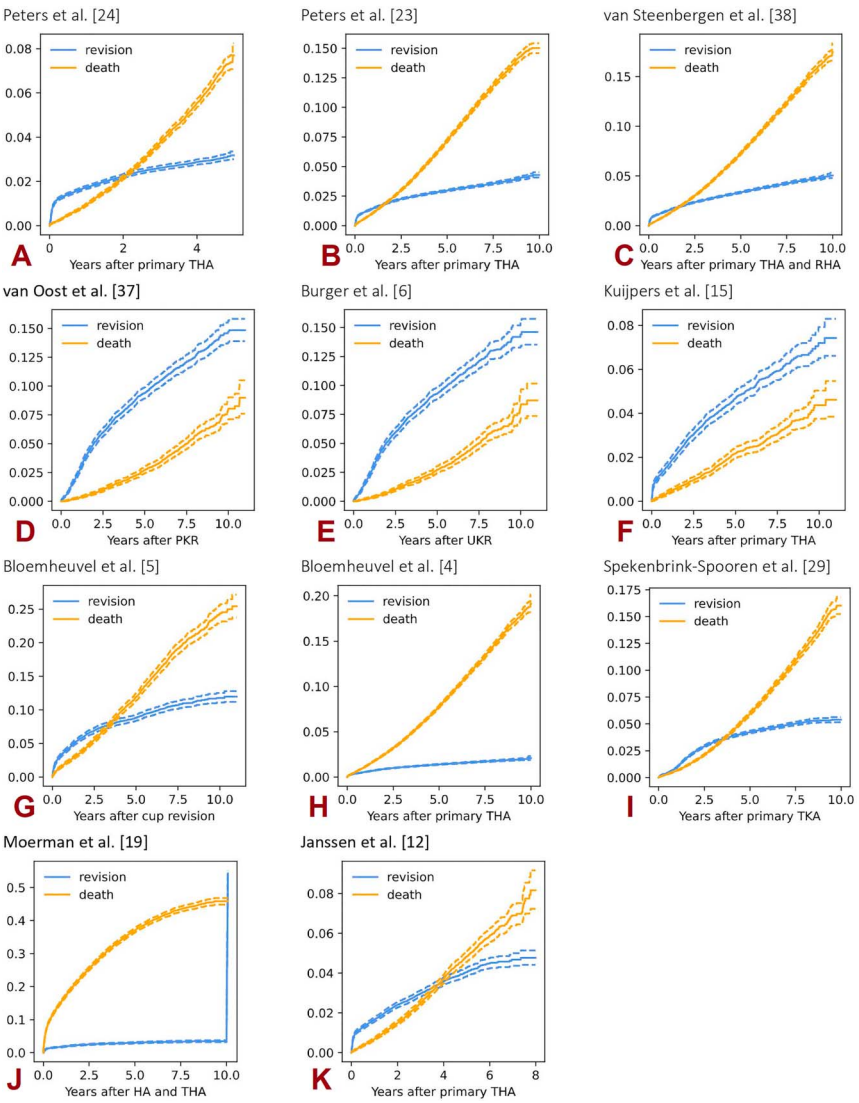

Fig. 3 These charts demonstrate the cumulative incidence function for the 11 datasets used in this study. Graph A = Aalen-Johansen curve for Peters et al. [24] for the event of revision and death after primary THA; Graph B = Aalen-Johansen curve for Peters et al. [23] for the event of revision and death after primary THA; Graph C = Aalen-Johansen curve for van Steenbergen et al. [38] for the event of revision and death after primary THA and RHA; Graph D = Aalen-Johansen curve for van Oost et al. [37] for the event of revision and death after PKR; Graph E = Aalen-Johansen curve for Burger et al. [6] for the event of revision and death after UKR; Graph F = Aalen-Johansen curve for Kuijpers et al. [15] for the event of revision and death after primary THA; Graph G = Aalen-Johansen curve for Bloemheuvel et al. [5] for the event of re-revision and death after cup revision surgery; Graph H = Aalen-Johansen curve for Bloemheuvel et al. [4] for the event of revision and death after primary THA; Graph I = Aalen-Johansen curve for Spekenbrink-Spooren et al. [29] for the event of revision and death after primary TKA; Graph J = Aalen-Johansen curve for Moerman et al. [19] for the event of revision and death after HA and THA; Graph K = Aalen-Johansen curve for Janssen et al. [12] for the event of revision and death after primary THA. The Aalen-Johansen curve plots the cumulative incidence function of the event of interest (revision) accounting for a competing risk (death). The x-axis represents the time after the index surgery (in years), the y-axis the cumulative incidence functions of revision and death. These curves provide insights into the probability of experiencing different types of events over time when multiple events (revision and death) are present. RHA = resurfacing hip arthroplasty; PKR = partial knee replacement; UKR = unicompartmental knee arthroplasty; HA = hemiarthroplasty. A color image accompanies the online version of this article.

This demonstrates that the conventional and machine-learning algorithms are insufficient for estimating the risk of revision for patients undergoing arthroplasty, with the current data available.

Discussion

Estimating the risk of revision after arthroplasty could inform patient and surgeon decision-making. However, there is a lack of well-performing prediction models assisting in this task, which may be owing to current conventional modeling approaches such as traditional survivorship estimators (Kaplan-Meier) or competing risk estimators (competing risk analyses). Recent advancements in machine learning survival analysis could improve decision support tools in this setting. In this comparative study, we found that a promising machine-learning approach (random survival forest) performed similarly to the traditional survivorship estimator. Neither machine-learning modeling nor traditional regression methods were sufficiently accurate in order to offer prognostic information in the clinical setting of predicting revision arthroplasty. The findings of this study suggest that the benefit of these modeling approaches may be limited in this context.

Limitations

First, the data were derived from the Dutch Arthroplasty Registry [9] and may not be generalizable to all registry populations. There is also the potential limitation of prediction modeling of revision after arthroplasty. Further research should validate these findings in geographically different settings, considering that other countries may collect varying sets of variables in their registry with longer follow-up durations. Second, we chose a common set of time-to-event points for a true comparison of model performances across the included datasets. Future studies should evaluate longer time-to-event points for individual studies investigating the benefits of machine learning survival analysis with a competing risk. Third, hyperparameter tuning was performed on the training dataset. We did not perform nested cross-validation because of the current computation time for training a random survival forest model. However, we did not expect to have an incremental benefit in model performance in our cohorts with the use of more sophisticated nested cross-validation.

Discussion of Key Findings

Our primary study goal was to estimate the model performance of machine learning compared with that of conventional modeling for estimating the likelihood of revision after arthroplasty in the presence of censored data. We included 11 datasets and separately compared the performance of machine learning to that of conventional modeling on those 11 datasets and found no incremental benefit to the use of machine-learning techniques.

Our findings were comparable to those of Aram et al. [2], who evaluated various model approaches for accurately estimating risk in patients undergoing revision after knee arthroplasty. Their results showed that a fully parametric model (random survival forest) is essential for predicting revision; however, their study concluded that such methods did not provide high discriminatory power at the individual level. Martin et al. [18] aimed to predict revision surgery after hip arthroscopy, including different model approaches (such as random survival forest), and concluded that there was limited clinical utility.

The finding that machine learning and traditional regression methods were comparable is consistent with a previous study from our group, which compared machine-learning and logistic regression algorithms for predicting binary outcomes in orthopaedic trauma using nine datasets [21]. In other fields, a study expected machine-learning analysis to outperform Cox proportional hazard regression analysis in breast cancer survival [20]. However again, random survival forest showed a similar performance to traditional regression analysis, and machine-learning algorithms that outperformed traditional regression analysis did not account for a competing risk.

These findings have implications for future research to improve decision support tools in the presence of competing risks. First, the observation that machine-learning models are comparable to traditional models in the presence of competing risks suggests that their benefit may be limited. Our findings highlight two points: Machine-learning methods should not be relied on heavily in prediction modeling, and the benefit of machine-learning models should be questioned for low-dimensional datasets. A low-dimensional dataset is a relatively small dataset with a manageable number of variables, and the specific threshold depends on the context and study. Most structured (tabular) orthopaedic datasets are considered low-dimensional datasets.

Second, the modeling approaches presented here are insufficient to predict the risk of revision after knee or hip arthroplasty. The low revision rate ranging between 0.5% and 4.6% may have limited the models' ability to distinguish between procedures with and without a revision in the current study's context [16]. Estimating the likelihood of revision in arthroplasty will likely remain challenging for this reason. Imbalance correction techniques could be applied before training the models in the future, but this comes at the cost of strong miscalibration [34]. Future research could compare machine learning and traditional

regression methods for other outcomes, such as patient-reported outcome measures, and evaluate patient satisfaction after arthroplasty [25, 28].

Conclusion

Neither machine-learning modeling nor traditional regression methods were sufficiently accurate to offer prognostic information in the clinical setting of predicting revision arthroplasty. The findings of this study suggest that the benefit of these modeling approaches may be limited in this context. Developing prediction models for estimating the risk of revision surgery in patients undergoing arthroplasty is challenging because of the censored nature of data and the current data availability. Future efforts should aim at validating this finding in other independent cohorts.

Group Authorship

Members of the Machine Learning Consortium include: Joost Colaris, Max Gordon, Prakash Jayakumar, Ruurd Jaarsma, Frank Ijpma, Zhibin Liao, Jasper Prijs, Lotte van der Linden, Johan Verjans, Mathieu Wijffels, and Yang Zhang.

Acknowledgments We thank the patients and staff contributing to the Dutch Arthroplasty Registry. In addition, we thank the Dutch Arthroplasty Registry for access to the selected research datasets.

References

1. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat*. 1978;5:141-150.
2. Aram P, Trela-Larsen L, Sayers A, et al. Estimating an individual's probability of revision surgery after knee replacement: a comparison of modeling approaches using a national data set. *Am J Epidemiol*. 2018;187:2252-2262.
3. Austin PC, Steyerberg EW, Putter H. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: cumulative total failure probability may exceed 1. *Stat Med*. 2021;40:4200-4212.
4. Bloemheuvel EM, van Steenbergen LN, Swierstra BA. Dual mobility cups in primary total hip arthroplasties: trend over time in use, patient characteristics, and mid-term revision in 3,038 cases in the Dutch Arthroplasty Register (2007-2016). *Acta Orthop*. 2019;90:11-14.
5. Bloemheuvel EM, van Steenbergen LN, Swierstra BA. Lower 5-year cup re-revision rate for dual mobility cups compared with unipolar cups: report of 15,922 cup revision cases in the Dutch Arthroplasty Register (2007-2016). *Acta Orthop*. 2019;90:338-341.
6. Burger JA, Kleeblad LJ, Sierevelt IN, et al. A comprehensive evaluation of lateral unicompartmental knee arthroplasty short to mid-term survivorship, and the effect of patient and implant characteristics: an analysis of data from the Dutch Arthroplasty Register. *J Arthroplasty*. 2020;35:1813-1818.
7. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Med*. 2015;13:1-10.
8. Cox DR. Two further applications of a model for binary regression. *Biometrika*. 1958;45:562-565
9. Dutch Arthroplasty Register. LROI Report 2020. Available at: https://www.lroi-report.nl/app/uploads/2022/04/PDF-LROI-annual-report-2021.pdf. Accessed July 4, 2022.
10. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94:496-509.
11. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics*. 2014;15:757-773.
12. Janssen L, Wijnands KAP, Janssen D, Janssen MWHE, Morrenhof JW. Do stem design and surgical approach influence early aseptic loosening in cementless THA? *Clin Orthop Relat Res*. 2018;476:1212-1220.
13. Keurentjes JC, Fiocco M, Schreurs BW, Pijls BG, Nouta KA, Nelissen RGHH. Revision surgery is overestimated in hip replacement. *Bone Joint Res*. 2012;1:258-262.
14. Kuijpers MFL, Hannink G, van Steenbergen LN, Schreurs BW. Outcome of revision hip arthroplasty in patients younger than 55 years: an analysis of 1,037 revisions in the Dutch Arthroplasty Register. *Acta Orthop*. 2020;91:165-170.
15. Kuijpers MFL, Hannink G, Vehmeijer SBW, van Steenbergen LN, Schreurs BW. The risk of revision after total hip arthroplasty in young patients depends on surgical approach, femoral head size and bearing type; an analysis of 19,682 operations in the Dutch arthroplasty register. *BMC Musculoskelet Disord*. 2019;20:385.
16. Labek G, Thaler M, Janda W, Agreiter M, Stöckl B. Revision rates after total joint replacement: cumulative results from worldwide joint register datasets. *J Bone Joint Surg Br*. 2011;93:293-297.
17. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18:e323.
18. Martin RK, Wastvedt S, Lange J, Pareek A, Wolfson J, Lund B. Limited clinical utility of a machine learning revision prediction model based on a national hip arthroscopy registry. *Knee Surg Sports Traumatol Arthrosc*. 2023;31:2079-2089
19. Moerman S, Mathijssen NMC, Tuinebreijer WE, Vochteloo AJH, Nelissen RGHH. Hemiarthroplasty and total hip arthroplasty in 30,830 patients with hip fractures: data from the Dutch Arthroplasty Register on revision and risk factors for revision. *Acta Orthop*. 2018;89:509-514.
20. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci Rep*. 2021;11:6968.
21. Oosterhoff JHF, Gravesteijn BY, Karhade AV, et al. Feasibility of machine learning and logistic regression algorithms to predict outcome in orthopaedic trauma surgery. *J Bone Joint Surg Am*. 2022;104:544-551.
22. Peters RM, van Steenbergen LN, Bulstra SK, et al. Nationwide review of mixed and non-mixed components from different manufacturers in total hip arthroplasty. *Acta Orthop*. 2016;87:356-362.
23. Peters RM, van Steenbergen LN, Stevens M, Rijk PC, Bulstra SK, Zijlstra WP. The effect of bearing type on the outcome of total hip arthroplasty. *Acta Orthop*. 2018;89:163-169.

24. Peters RM, van Steenbergen LN, Stewart RE, et al. Patient characteristics influence revision rate of total hip arthroplasty: American Society of Anesthesiologists score and body mass index were the strongest predictors for short-term revision after primary total hip arthroplasty. *J Arthroplasty*. 2020;35:188-192.e2.
25. Pickett KL, Suresh K, Campbell KR, Davis S, Juarez-Colunga E. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC Med. Res Methodol.* 2021;21:216.
26. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26: 2389–2430.
27. Rubin D. *Multiple Imputation for nonresponse in surveys*. John Wiley & Sons Inc; 1987.
28. Sorel JC, Veltman ES, Honig A, Poolman RW. The influence of preoperative psychological distress on pain and function after total knee arthroplasty: a systematic review and meta-analysis. *Bone Joint J*. 2019;101:7-14.
29. Spekenbrink-Spooren A, Van Steenbergen LN, Denissen GAW, Swierstra BA, Poolman RW, Nelissen RGHH. Higher mid-term revision rates of posterior stabilized compared with cruciate retaining total knee arthroplasties: 133,841 cemented arthroplasties for osteoarthritis in the Netherlands in 2007-2016. *Acta Orthop*. 2018;89:640-645.
30. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur. Heart J.* 2014;35:1925-1931.
31. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128-138.
32. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Software Artic*. 2011; 45:1-67.
33. van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making*. 2015;35:162-169.
34. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Informatics Assoc*. 2022;29:1525-1534.
35. van der Pas S, Nelissen R, Fiocco M. Different competing risks models for different questions may give similar results in arthroplasty registers in the presence of few events. *Acta Orthop*. 2018;89:145-151.
36. van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. *BMJ*. 2022;377:e069249.
37. van Oost I, Koenraadt KLM, van Steenbergen LN, Bolder SBT, van Geenen RCI. Higher risk of revision for partial knee replacements in low absolute volume hospitals: data from 18,134 partial knee replacements in the Dutch Arthroplasty Register. *Acta Orthop*. 2020;91:426-432.
38. van Steenbergen L, Denissen G, Schreurs B, Zijlstra W, Koot H, Nelissen R. Dutch advice not to use large head metal-on-metal hip arthroplasties justifiable – results from the Dutch Arthroplasty Register. *Ned Tijdschr voor Orthop*. 2020;27:4-11.
39. Zijlstra WP, De Hartog B, Van Steenbergen LN, Scheurs BW, Nelissen RGHH. Effect of femoral head size and surgical approach on risk of revision for dislocation after total hip arthroplasty. *Acta Orthop*. 2017;88:395-401.